ABSTRACT OF THE DISCLOSURE

A context automaton such as a left context automaton predefined and a right context automaton generate a context record that is combined with pattern knowledge stored in a token automaton to segment an input data stream into tokens. The resulting context-aware tokenizer can be used in many natural language processing application including text-to-speech synthesizers and text processors. The tokenizer is robust in that upon failure to match any explicitly stored token pattern a default token is recognized. Token matching follows a left-to-right longest-match strategy. The overall process operates in linear time, allowing for fast context-dependent tokenization in practice.